The @Note Biomedical Text Mining Workbench

Anália Lourenço[1], Rafael Carreira[2], Sónia Carneiro[1], Paulo Maia[2], Eugénio C. Ferreira[1], Miguel Rocha[2], Isabel Rocha[1]

[1]IBB - Institute for Biotechnology and Bioengineering, Center of Biological Engineering

[2]Departament of Informatics / CCTC

University of Minho

Campus de Gualtar, 4710-057 Braga - PORTUGAL

Abstract.

Biomedical scientific publishing grows at a steady rate and research goals are becoming ever more focused and complex. The ability to link structured database information to the essentially unstructured scientific literature and to extract additional information is invaluable.

Biomedical Text Mining, i.e., the field that deals with the automatic retrieval and processing of biomedical literature, is perhaps one of today's most promising research fields. The automated processing of free text is not exactly a new computational issue. Text Mining has a long experience in the retrieval and processing of general text. The development of search engines and indexed directories, the compilation of dictionaries and the automatic translation of documents have impelled the development of powerful text processing techniques. Nevertheless, biomedical texts present additional challenges: the terminology does not always follows standard nomenclatures; new terms are constantly emerging; and term homonymy and synonymy (including term variants and abbreviations) make it very difficult to accurately identify entities.

The @Note Biomedical Text Mining workbench represents a commitment between practical use and research, integrating current Biomedical Text Mining methods into an extensible workbench and providing biologists with intuitive tools capable of supporting their bibliographic searches and further literature

curation. Its design directives are two-fold: flexibility and interoperability towards the inclusion and further extension of state-of-the-art Biomedical Text Mining approaches; transparency and simplicity, enabling the use of techniques without requiring expert knowledge about the undergoing activities.

Currently, @Note supports PubMed search for relevant documents and document retrieval from open-access and subscribed Web-accessible journals. Automatic document annotation encompasses the recognition of several classes of biological entities based on lexicon support and the extraction of relationships for known biology-related verbs. Additionally, manual curation can refine document annotation and enhance lexicon support.

The workbench is available at http://sysbio.di.uminho.pt/aNote.php